Advanced Topics in Artificial Intelligence - (Task I)
Deadline: TBF

A bayesian network represents a joint network distribution over a number of random variables (RVs). In this task, we ask you to implement inference for such models, and give two alternatives.

Introduction

The BN representation a joint distribution between RVs. BNs simplify full table representation by taking advantage of independence between RVs. We shall assume discrete *boolean* RVs next, but most techniques will work well for any discrete networks (but not for contiguous or hybrid).

Our goal is to answer queries of the form $Pr(A_i|B_1 \dots B_n)$, where $B_1 \dots, B_n$ are the *evidence* variables, that is, variables for which we observed a certain value, and A_i is an unobserved RV, the marginal. We want to find out the probability of the marginal after observing the evidence.

The probability of a joint distribution with N RVs is $Pr(X_1 = x_1, \dots, X_N = x_n)$ for a specific value of $X_1, \dots, X_N$. Summing over all possible cases:

$$\sum_{X_1} \dots \sum_{X_N} Pr(X_1 = x_1), \dots, X_N = x_N) = 1$$

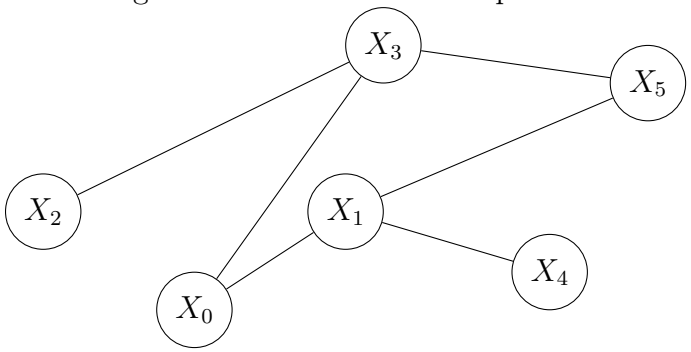
A joint probability can be always converted into a product of factors by using the famous chain rule:

$$Pr(X_1|X_2 \dots X_1) \dots, X_N) \dots Pr(X_{N-1}|X_N)Pr(X_N)$$

A bayesian network is a graphical depiction of a joint distribution, where:

- we assume an ordering between RVs, and use this ordering to generate the graph;
- for each $Pr(X_i|X_{i+1} \dots X_N)$ we drop all X_k such that $Pr(X_i|X_j \dots X_m) = Pr(X_i|X_k X_j \dots X_m)$. The remaining variables are called X_i's parents;
- we draw edges from parents to children.

The figure below shows an example of AaBN with 6 RVs.



The chain formula is:

$$\sum_{X_0,X_1,X_2,X_3,X_4,X_5} Pr(X_0|X_1X_3)Pr(X_1|X_5)Pr(X_2|X_3)Pr(X_3|X_5)11Pr(X_4|X_1)Pr(X_5)$$

To cope with evidence, say $X_1 = \mathtt{t}$, we just drop summands that requuire $\neg V_1$:

$$\sum_{X_0,X_2,X_3,X_4,X_5} Pr(X_0|x_1X_3)Pr(x_1|X_5)Pr(X_2|X_3)Pr(X_3|X_5)Pr(X_4|x_1)Pr(X_5)$$

Notice that setting $X_1 = \mathtt{t}$ halves the number the terms in the summation. Next, to compute a marginal probability for X_3, say $Pr(X_3|x_1)$, we can use the definition of conditional probability to obtain the two cases:

$$Pr(x_3|x_1) = \frac{Pr(x_3,x_1)}{Pr(x_1)} = \frac{\sum_{X_0,X_2,X_4,X_5} Pr(X_0|x_1x_3)Pr(x_1|X_5)Pr(X_2|x_3)Pr(x_3|X_5)Pr(X_4|x_1)Pr(X_5)}{\sum_{X_0,X_2,X_3,X_4,X_5} Pr(X_0|x_1X_3)Pr(x_1|X_5)Pr(X_2|X_3)Pr(X_3|X_5)Pr(X_4|x_1)Pr(X_5)}$$

Normalization If the variables are boolean, we always have two cases, and only two: $Pr(\overline{x_3}|x_1) + Pr(x_3|x_1) = 1$. Also:

$$Pr(\overline{x_3}|x_1) = \frac{Pr(\overline{x_3},x_1)}{Pr(x_1)} = \frac{\sum_{X_0,X_2,X_4,X_5} Pr(X_0|x_1\overline{x_3})Pr(x_1|X_5)Pr(X_2|\overline{x_3})Pr(\overline{x_3}|X_5)Pr(X_4|x_1)Pr(X_5)}{\sum_{X_0,X_2,X_3,X_4,X_5} Pr(X_0|x_1X_3)Pr(x_1|X_5)Pr(X_2|X_3)Pr(X_3|X_5)Pr(X_4|x_1)Pr(X_5)}$$

which is very similar to the formula for $Pr(x_3|x_1)$.
We thus get $\frac{Pr(x_3,x_1)}{Pr(x_1)} + \frac{Pr(\overline{x_3},x_1)}{Pr(x_1)} = 1$, hence

$$Pr(x_3|x_1) = \frac{Pr(x_3,x_1)}{Pr(x_3,x_1) + Pr(\overline{x_3},x_1)}$$

Notice that we can ask questions about any variable in the network. Intuitively, questions often are:

1. given evidence on root nodes, how do the probabilities for the observed variables change? An example would be: does forcing X_5 to always true makes any difference on X_2?

2. given that X_0 is true, does our expectation that $X_2 = \mathtt{t}$ increases?

These are a few basic concepts of BNs. There are several excellent BN libraries and repositories of BNs. Examples include the Python packages https://pomegranate.readthedocs.io/en/latest/ and .

We expect you to use existing Input-Output code and existing BNs in this project, and focus on the algorithms and their limitations.

Alternative I: Gibbs Sampling.

This method is based on MarkovChain Monte Carlo, MCMC. Starting from an initial sample, we keep on generating samples according to the BN distribution, and stop when our estimate of the marginal is stable. The algorithm is:

1. Initialize the RVs with a random set of values and store them in a vector, say: $V = \{0, 1, 1, 0, 0, 0\}$ your assignment must agree with existing evidence.
2. While not converging, loop through every RV:
   (a) for every value x_i in the RV X_i estimate $Pr(x_i| \ldots, X_{i-1} = V_{i-1} X_{i+1} = V_{i-+} \ldots)$.
   (b) sample a new v_i from the x_i.
3. generate the estimate for $Pr(x_5| \ldots)$; the estimator for X_i is just the number of times $X_i = t$ over the number of iterations.

Key Issues

To implement the algorithm we need to compute $Pr(x_i| \ldots, X_{i-1} = V_{i-1} X_{i+1} = V_{i+1} \ldots)$. Using normalization we have:

$$Pr(x_5| \ldots) = \frac{Pr(x_5, \ldots)}{Pr(x_5, \ldots) + Pr(\overline{x_5}, \ldots)}$$

This sounds scary, but can be simplified, Imagine the sample is $\{1, 1, 1, 1, 1, 1, 1\}$ and we want to process X_5. Our formula is:

$$\frac{Pr(x_0|x_1x_3)Pr(x_1|x_5)Pr(x_2|x_3)Pr(x_3|x_5)Pr(x_4|x_1)Pr(x_5)}{Pr(x_0|x_1x_3)Pr(x_1|x_5)Pr(x_2|x_3)Pr(x_3|x_5)Pr(x_4|x_1)Pr(x_5) + Pr(x_0|x_1x_3)Pr(x_1|\overline{x_5})Pr(x_2|x_3)Pr(x_3|\overline{x_5})Pr($$

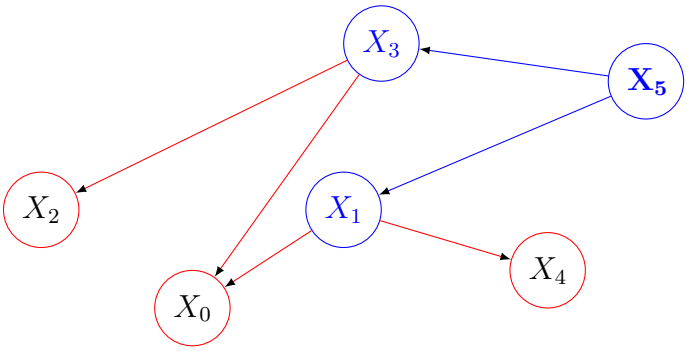
The three CPTs $Pr(X_0|X_1X_3)$, $Pr(X_2|X_3)$, and $Pr(X_4|X_1)$ are independent of the value of X_5. Reordering the formula gives

$$\frac{Pr(x_0|x_1x_3)Pr(x_2|x_3)Pr(x_4|x_1)Pr(x_1|x_5)Pr(x_3|x_5)Pr(x_5)}{Pr(x_0|x_1x_3)Pr(x_2|x_3)Pr(x_4|x_1)Pr(x_1|x_5)Pr(x_3|x_5)Pr(x_5) + Pr(x_0|x_1x_3)Pr(x_2|x_3)Pr(x_4|x_1)Pr(x_1|\overline{x_5})Pr($$

the red terms can be factored out: this gives

$$\frac{Pr(x_1|x_5)Pr(x_3|x_5)Pr(x_5)}{Pr(x_1|x_5)Pr(x_3|x_5)Pr(x_5) + Pr(x_1|\overline{x_5})Pr(x_3|\overline{x_5})Pr(\overline{x_5})}$$

or graphically:



Hence, for each variable X_i, we just need to multiply the probabilities in the factors X_i that include X_i. The value can be precomputed. This set of variables is known as the Markov Blanket, $\mathcal{M}(\mathcal{X}_)$: it is the smallest set that, if instantiated, determines the probability of the RV X_i. In other words, if we have complete evidence on $\mathcal{M}(X_i)$ and $\mathcal{M}(X_i) \subset RVs$, then $Pr(X|Vs) = Pr(X_i|\mathcal{M}(X_i))$.

Implementation

The one remaining problem is how to detect convergence (the algorithm is *not* guaranteed to converge). A sequence of steps is called a chain. We can stop the process if:

- we exceeded a maximum number of steps;
- the maximum/average change in probabilities is under a threshold.

Last, two important techniques are:

- warm-up: drop the first 100-1000 iterations, as they may depend too much on the initial guess;
- multiple chains: just run the algorithm several times, using independent samples. If all the chains converge for the same estimates, you are most likely converging.

Variable Elimination

The other algorithm we consider here is Variable Elimination (VE). The idea is simply to remove RVs until only the marginal remains. This is straightforward for evidence, but can be expensive for other variables.

The algorithm is:

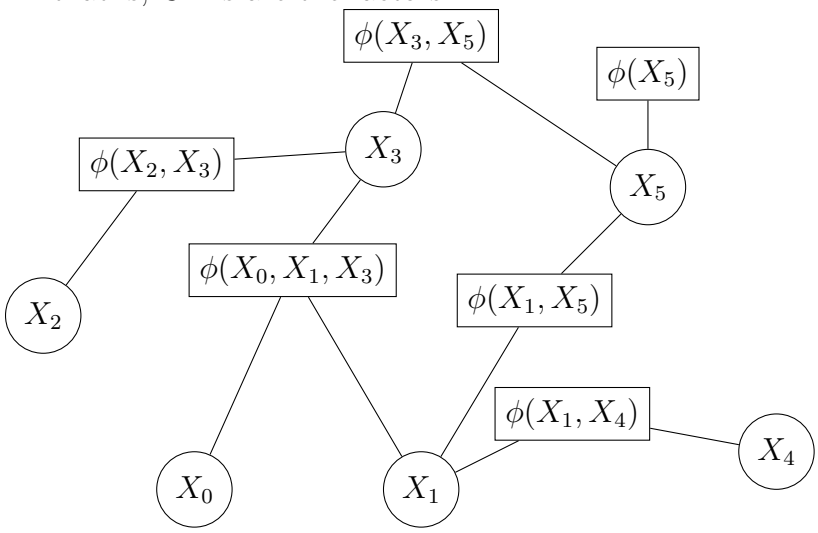
1. for every VE X_i with evidence x_i, assign all entries $Pr(X_i = v, v \neq x_i)$ must be set to 0, and $Pr(X_i = x_i) = 1$.
2. eliminate every other remaining variable, one by one.
3. normalize the final result.

0.1 Factors

Throughout the process, we shall multiply and sum conditional probabilities. The results will be numbers, but we will only have probabilities at the very end, when we normalise. As intermediate terms will be part of a product, we shall call them *factors*. Initially,

$$\phi(X_i, X_j, X_k) = Pr(X_i|X_jX_k)$$

that is, CPTs are the factors.

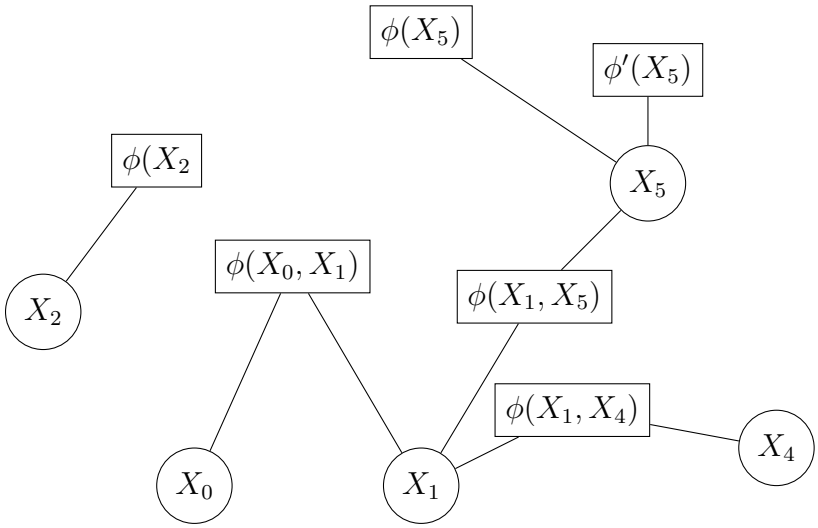


The graph above shows the network in terms of factors. Notice that factors are undirected: they just relate variables. Every variable must belong to a factor: there are three variables that belong to a single factor, and three variables that belong to 3 factors. Notice also that we have a factor with a single variable, $\phi(X_5)$.

0.2 Complete Example

Let us estimate $Pr(X_4|x_3)$, that is, how do probabilities change if $X_3 = \texttt{t}$

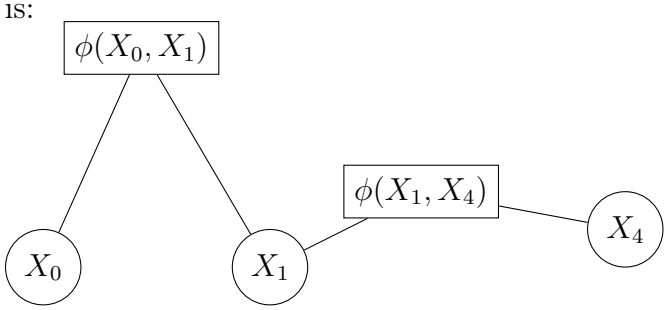
First step, we "purge" X_3: we simply remove all entries such that $X_3 = \texttt{f}$. The remaining entries will have $X_3 = \texttt{t}$, so we can just drop X_3 of the factor. We get:



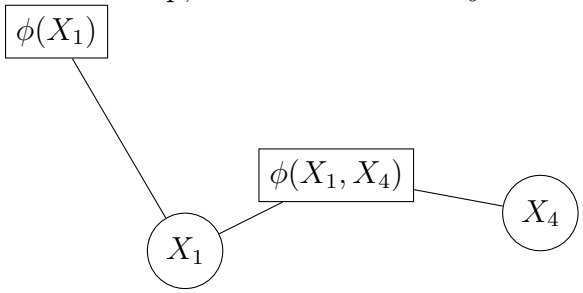
Notice X_2 is now independent of the remaining variables: we can forget about it. We are left with X_0, X_1 and X_5. Eliminating X_0 corresponds to merging all factors that share X_0 into a single X_0 and the same for X_1, X_5. In detail:

- X_0 is in a single factor with X_1, so the result will be a factor $\phi(X_1)$;
- X_1 is in three factor with all remaing variables, so the result will cover all remaining variables.
- X_5 is in three factors, but only shares with X_5.

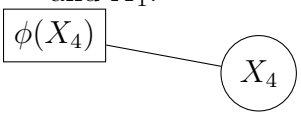
Using the smaller factors first heuristic, X_0 and X_5 are equally best. Choosing X_5 we need to multiply $\phi(X_1, X_5)$, $\phi'(X_5)$, $\phi(X_5)$.After doing that we can project on X_1, and the result is:



Next step, we can eliminate X_0:



and X_1:



We can now use the factor $\phi(4) = [\alpha, \beta]$ to compute our probability distribution as $Pr(X_4|X_3 = \mathtt{t}) =$

$$\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}$$

0.3 The Algorithm

Given a graph G, evidence EV, and query Q:

1. : Initialise
   (a) propagate evidence, as described above ;
   (b) $RV = VARS(G)$
   (c) for every V add a factor $\phi_V(...)$
   (d) for every V compute the size of the factor generated by eliminating V, amd only V. Order these numbers in a queue.
2. while $RV \neq \{EV\}$:
   (a) eliminate best G.
   (b) add new factor.
3. Normalise.

To eliminate V:

- colllect all factors with V, $\sum$.
- Let **rv** be the ser of all RVs in $\sum$:
- replicate every factor until all have the same size;
- multiply factors element by element;
- sum out V by adding the true and false cases.

To understand elimination, let us study the first step of the example in detail. we have the following product:

$$\phi(X_0, X_1)\phi(X_1|X_5)\phi(X_2)\phi(X_5)\phi(X_4|X_1)\phi(X_5)$$

The product must be summed over all the values of the unknown variables.

$$\sum_{X_1,X_2,X_4,X_5} \phi(X_0,X_1)\phi(X_1|X_5)\phi(X_2)\phi(X_5)\phi(X_4|X_1)\phi(X_5)$$

We can separate the summands into a sum for X_2 and the sum for the rest. We also can move the factors left or right, so we have:

$$\sum_{X_1,X_2,X_4,X_5} \sum_{X_2} \phi(X_2)\phi(X_0,X_1)\phi(X_1|X_5)\phi(X_5)\phi(X_4|X_1)\phi(X_5)$$

and apply distributivity:

$$(\sum_{X_2} \phi(X_2))(\sum_{X_1,X_2,X_4,X_5} \phi(X_0,X_1)\phi(X_1|X_5)\phi(X_5)\phi(X_4|X_1)\phi(X_5))$$

We can see $(\sum_{X_2} \phi(X_2))$ as a constant, and let it multiply the denominator Z. Normalization will deal it.

We would like to do something similar with X_5. In this case, we would send the $\phi(\dots X_5)$ to the left and spllt X_5:

$$\sum_{X_1,X_2,X_4} \sum_{X_5} \phi(X_1|X_5)\phi(X_5)\phi(X_5)\phi(X_0,X_1)\phi(X_4|X_1))$$

The first three factors depend on X_5, but the last two do not, so we can:

$$\sum_{X_1,X_2,X_4} \phi(X_0,X_1)\phi(X_4|X_1) \sum_{X_5} \phi(X_1|X_5)\phi(X_5)\phi(X_5))$$

Multiplying and adding over X_5 we get:

$$\sum_{X_1,X_2,X_4} \phi(X_0,X_1)\phi(X_4|X_1)\phi(X_1)$$

To multiply, we need to consider every possible case of X_5 and X_1: we need to have the same dimensions and then we just aply the Hadamard product. X_5 is removed by summing the two cases.